REMARKS

This response is filed in reply to the Office Action mailed August 18, 2003. Claims 1-4, 8 and 9 have been canceled by the above amendment and new claims 24-50 have been added. Support for the new claims can be found throughout the specification. Specifically, support for claims 24-30 can be found at page 5, lines 1-29. Support for claims 31-33 can be found at page 12, lines 29-31, bridging to page 13, lines 1-18. Support for claims 36-38 can be found at page 14, lines 21-25. Support for new claims 39-50 can be found in Example 10 (see at page 30, lines 1-19).

Applicants note that non-elected claims 14-23 have been withdrawn from consideration. Applicants will request that withdrawn claims 17-23 be rejoined pursuant to MPEP §821.04 once the elected claims are deemed allowable.

No new matter has been added. Claims 24-50 are pending and at issue. Applicants request reconsideration of the pending claims.

INTERVIEW SUMMARY

A telephonic interview between the Examiner and Applicants representative took place on July 10, 2003. Applicants representative requested that a substitute office action be provided because the Gonzalez reference had not been included in the original office action. Applicants representative further requested that the time for responding to the office action be reset.

OBJECTION

Claim 8 is objected to as being dependent upon a non-elected base claim. This objection is moot in view of the cancellation of claim 8.

I. REJECTION UNDER 35 U.S.C. §101

Claims 1-4, 8 and 9 stand rejected under 35 U.S.C. §101 as allegedly directed to non-statutory subject matter. This rejection is moot in view of the cancellation of claims 1-4, 8 and 9.

Applicants note that the new claims now recite an "isolated" polypeptide, clearly distinguishing the claimed compositions from products of nature.

II. REJECTIONS UNDER 35 U.S.C. §112, FIRST PARAGRAPH

Written Description

Claims 1-3 and 8 stand rejected under 35 U.S.C. §112, first paragraph, as allegedly containing subject matter not described in the specification in such a way as to reasonably convey to one skilled in the relevant art that the inventor, at the time the application was filed, had possession of the claimed invention. This rejection is moot in view of the cancellation of claims 1-3 and 8. Applicants traverse this rejection as it may apply to the new claims.

The Office Action states that the claims are drawn to a "genus of (R)-2,3-dehydrogenase, with any structure and from any source." Applicants submit that one skilled in the art would not expect the genus of proteins encompassed by the pending claims to have substantial variation. The new claims encompass a narrow range of variants including those (a) having up to 50 conservative amino acid substitutions, (b) hybridizing under high stringency conditions with the nucleic acid of SEQ ID NO: 1, or (c) having at least 70% identity to the amino acid sequence of SEQ ID NO: 2. In addition, many of the pending claims are substantially narrower than those. As there is little variation within the claimed genus, a single example is sufficient to demonstrate possession of the claimed genus of polypeptides.

For example, claim 24 recites an isolated polypeptide comprising a sequence with "70%" sequence identity to the polypeptide of SEQ ID NO:2. The polypeptides encompassed by claim 24 are further limited to butanediol dehydrogenases possessing specific functional characteristics recited in parts (a), (b) and (c) of claim 24. Any such proteins would necessarily be significantly similar in terms of structure. Similar arguments are applicable to the polypeptides claimed in claim 26. Dependent claims 25, 27 and 31-33 narrow the scope even more.

Claim 28 recites specific chemical and physical properties such as enzymatic activity and molecular weight. The characteristics recited in claim 28 clearly limit the claimed genus of polypeptides to those having very similar structure and function, i.e., they are limited to a

butanediol dehydrogenase with the specific chartacteristics set forth in the claim. The "Guidelines for Examination of Patent Applications Under the 35 USC §112, Written Description Requirement" (herein after "Guidelines") support claiming in this manner. The Guidelines indicate that patent examiners are required to determine "what each claim, as a whole, covers" (66 Fed. Reg. 1099, at 1105). The Guidelines further indicate that the disclosure of any combination of identifying characteristics that "distinguish the claimed invention from other materials and would lead one of skill in the art to the conclusion that the applicant was in possession of the claimed species" is sufficient to comply with the written description requirement (Id at 1106).

Claims 39, 42, 45 and 48 recite individual fragments, or a combination of fragments, of the polypeptide set forth in SEQ ID NO:2. For example, claim 39 recites a polypeptide that is a (R)-2,3-butanediol dehydrogenase and comprises the amino acid sequence of SEQ ID NO:5 (Phe-His-Ala-Ala-Phe-Asp). Claim 39 further recites functional characteristics that the claimed dehydrogenase must possess. Applicants submit that the fragments disclosed in the originally filed specification provide "blaze marks" (In re Ruschig, 379 F.2d 990, 994, (CCPA 1967)) directing the skilled artisan to polypeptides possessing the required activities of the claimed polypeptides. A person of skill in the art would not expect substantial variation among species encompassed within the scope of the claims. Similar arguments can be made for claim 42 (SEQ ID NO:4; Ala-Thr-Ser-His-Cys-Ser-Asp-Arg-Ser-Arg-Tyr-Lys-Asp-Thr-Val-Ala-Gln-Asp-Leu-Gly-Leu), claim 45 (SEQ ID NO:3; Lys-Pro-Gly-Asp-Arg-Val-Ala-Val-Glu-Ala) and claim 48 (SEQ ID NOs:3, 4 and 5).

In view of the limitations recited in the pending claims, the skilled artisan would recognize that the inventor was in possession of the claimed polypeptides as of the filing date of the application.

Enablement

Claims 1-3 and 8 also stand rejected under §112, first paragraph, as allegedly containing subject matter not described in the specification in such a way as to enable one of skill in the art to make or use the invention. This rejection is moot in view of the cancellation of claims 1-3 and 8. Applicants traverse this rejection as it may apply to the new claims.

The Office Action alleges that the specification does not reasonably provide enablement for an (R)-2,3-butanediol dehydrogenase different from SEQ ID NO:2. Applicants respectfully disagree. As noted above, new claim 24 claims a polypeptide having at least 70% sequence identity to the amino acid sequence set forth in SEQ ID NO:2. In addition, the claimed polypeptide is limited to a "butanediol dehydrogenase" having the functional characteristics recited in parts (a), (b) and (c) of claim 24. With regard to the functional limitations set forth in parts (a) and (b), the specification provides an assay for identifying oxidizing (2R,3R)-2,3-butanediol activity recited in part (a) (see page 8, lines 6-12) and an assay for identifying activity of oxidizing glycerol recited in part (b) (page 8, lines 13-18). The specific activity of the enzyme recited in part (c) can be calculated from the previously described assays. Producing a polypeptide that comprises a sequence at least 70% identical to SEQ ID NO:2 is a trivial task that any molecular biologist could accomplish using standard recombinant DNA techniques. Applicants have taught how to assay for the indicated activities. Any polypeptide that possesses the activity specified in the claim is presumptively useful as a (R)-2,3-butanediol dehydrogenase. Thus, the specification teaches both how to make and how to use the claimed polypeptides. The enablement requirement is clearly met.

Similar arguments can be made for pending claims 26, 28, 39, 42, 45 and 48. One skilled in the art could make the claimed polypeptides from the disclosures in the patent coupled with information known in the art without undue experimentation. By providing the appropriate nucleic acid and amino acid sequences, along with additional information regarding sequence comparison programs, hybridization conditions, and enzymatic assays, Applicants have presented the skilled artisan with all the information necessary to make the claimed compositions using only routine experimentation. Once made, the polypeptides can be used as described in the specification.

On page 6, the Office Action asserts that the specification fails to establish the specific structures responsible for the butanediol dehydrogenase activity of the claimed polypeptide or what changes in the structures could be made without abolishing this activity (see page 6, parts (A) through (D)). The Office Action concludes that undue experimentation would be required to

enable the full scope of the pending claims because neither the specification nor the art provides any guidance as to which of the polypeptides encompassed by the claims will retain the activity of the polypeptide set forth in SEQ ID NO:2 (i.e., butanediol dehydrogenase activity).

Applicants first note that, because the claim requires that the polypeptide possess butanediol dehydrogenase activity, every polypeptide encompassed by the claims necessarily retains that activity. Thus, encompassing inactive proteins is simply not an issue.

Second, Applicants agree that it is possible, at least in some cases, to abolish activity of a given protein by mutating a critical residue. However, Applicants disagree that this fact means that one of ordinary skill cannot make functional analogs of the claimed polypeptide without undue experimentation. In support of this, Applicants refer to Bowie et al. (*Science* **247**:1306; copy enclosed as Appendix A) which teaches, at page 1306, col.2, lines 12-13, that "proteins are surprisingly tolerant of amino acid substitutions." Bowie et al. cites as evidence a study carried out on the lac repressor. Of approximately 1500 single amino acid substitutions at 142 positions in this protein, about one-half of the substitutions were found to be "phenotypically silent": that is, had no noticeable effect on the activity of the protein (page 1306, col. 2, lines 14-17). Presumably the other half of the substitutions exhibited effects ranging from slight to complete abolishment of repressor activity. Thus, one can expect, based on Bowie et al.'s teachings, to find over half (and possibly well over half) of random substitutions in any given protein to result in mutated proteins with full or nearly full activity. These are far better odds than those at issue in In re Wands, 858 F.2d 731 (Fed. Cir. 1988), in which the court said that screening many hybridomas to find the few that fell within the claims was not undue experimentation. The question is not whether it is possible to abolish activity with a single amino acid change, but rather whether one of ordinary skill can produce, without undue experimentation, mutants in which the activity is not abolished. Based on Bowie et al.'s teachings, one would predict that even random substitution of residues in the claimed polypeptide will result in a majority of the mutants' having full or partial butanediol dehydrogenase activity.

Furthermore, the specification amply teaches how to make and test mutants to find those with the dehydrogenase activity required by the claims. For example, there is little doubt that the

provision of an amino acid sequence (e.g., SEQ ID NO:2) necessarily provides a person of average skill in the art with enough information to enable him or her to make a sequence containing, for example, up to 50, 30, or 10 conservative amino acid substitutions. Further, generation of such mutants is a routine task. Similarly, there is little doubt that the provision of the amino acid sequence of a fragment (e.g., SEQ ID NO:3, 4 and 5) necessarily provides a person of average skill in the art with enough information to enable him or her to make a butanediol dehydrogenase containing that sequence by simply following the teachings of the specification. Plainly, these experiments would not be "undue."

In view of the limitations recited in the pending claims, the skilled artisan could make the claimed polypeptides from the disclosures in the patent coupled with information known in the art without undue experimentation.

III. REJECTIONS UNDER 35 U.S.C. §112, SECOND PARAGRAPH

Claims 1-3 and 8 stand rejected under 35 U.S.C. §112, second paragraph, as allegedly indefinite for failing to particularly point out and distinctly claim the subject matter which the Applicants regard as the invention. This rejection is moot in view of the cancellation of claims 1-3 and 8. Applicants note that the new claims do not employ the language that the office action describes as confusing.

IV. REJECTION UNDER 35 U.S.C. §102

§102(a)

Claims 1 and 8 stand rejected under 35 U.S.C. §102(a) as allegedly anticipated by Gonzalez et al. Applicants traverse this rejection to the extent it may be applied to the presently pending claims, but also submits herewith a certified copy of a translation of the Japanese patent application to which the pending application claims priority. The filing date of the priority document is October 31, 2000. The publication date of the cite reference of Gonzalez is November 17, 2000. Accordingly, the cited reference cannot anticipate the pending claims.

§102(b)

Claims 1 and 8 stand rejected under 35 U.S.C. §102(b) as allegedly anticipated by Heidlas et al. This rejection is moot in view of the cancellation of claims 1 and 8. Applicants traverse this rejection as it may be applied to the new claims.

The dehydrogenase disclosed in Heidlas was purified from *Saccharomyces cerevisiae*. While Heidlas fails to provide any nucleic acid or amino acid sequence information associated with their purified dehydrogenase, the cited reference of Gonzalez does disclose the sequence of a dehydrogenase derived from *Saccharomyces cerevisiae*. Applicants contend that the dehydrogenase of Gonzalez is the same as the dehydrogenase disclosed in Heidlas. Appendix B, part 1, which accompanies the present response, provides a comparison of the amino acid sequence set forth in SEQ ID NO:2 (derived from the genus *Pichia*) with the amino acid sequence disclosed in Gonzalez (derived from the genus *Saccharomyces*). The sequence comparison clearly indicates that the two sequence share only about 48% homology. In contrast, the subject matter of claim 24 is limited to those polypeptides having "70%" sequence homology to SEQ ID NO:2. Claims 34 and 35 claim polypeptides that "consist of" or "comprise" the amino acid sequence set forth in SEQ ID NO:2. The subject matter of claims 36-38 is limited to polypeptides that differ from SEQ ID NO:2 by, at most, 13% (i.e., 50 amino acid substitutions). Thus, the claimed polypeptides would still be required to have at least 87% sequence homology with SEQ ID NO:2. Accordingly, the subject matter of the previously mentioned claims are not anticipated by the cited reference of Heidlas.

New claims 39, 42, 45 and 48 recite SEQ ID NO:5, SEQ ID NO:4, SEQ ID NO:3 or SEQ ID NOs:5, 4 and 3, respectively. These amino acid sequences are tryptic peptide fragments of SEQ ID NO:2. The sequence comparison provided in Appendix B, part 1, clearly indicates that none of the fragments exist in the sequence set forth in the cited reference. Thus, polypeptides of claims 39, 42, 45 and 48 can not be anticipated by the cited reference.

Claim 26 claims a polypeptide encoded by a polynucleotide that is at least 80% identical to a polynucleotide comprising the nucleotide sequence of SEQ ID NO:1. Applicants submit that the amino acid sequence data provided in Appendix B, part 1, strongly indicates

that the cited reference fails to teach a nucleic acid sequence with the limitations recited in claim 26, or a polypeptide encoded thereby. Accordingly, the cited reference fails to anticipate the subject matter of claim 26.

Claim 28 is limited to a dehydrogenase that ... "has a molecular weight of about 36,000 Da when determined by sodium dodecyl sulfate-polyacrylamide gel electrophoresis and about 76,000 Da when determined by gel filtration." At page 269, column 2, under the heading "Molecular mass," Heidlas discloses a dehydrogenase having a molecular weight of about "140,000" by gel filtration, clearly distinguishing the dehydrogenase of the cited reference from the dehydrogenase of claim 28.

In view of the new claims and in light of the above discussion, Applicants submit that the cited reference fails to anticipate each and every element of the claimed invention.

V. PROVISIONAL OBVIOUSNESS-TYPE DOUBLE PATENTING

Claims 1 and 8 stand provisionally rejected under the judicially created doctrine of obviousness-type double patenting as allegedly unpatentable over claims 1 and 3-4 of co-pending Application No. 10/147,003 ('003). Applicants believe that this rejection is moot in view of the cancellation of claims 1 and 8. It is further believed that the claims pending in the present application are patentably distinct from those of the '003 co-pending application because the sequences (both nucleic acid and amino acid) claimed in the present application do not encompass the sequences claimed in the '003 application. For example, the homology between SEQ ID NO:1 of the present invention and the SEQ ID NO:1 of the '003 application (see page 16 of the '003 specification) is only about 4% (see Appendix B, part 2). The homology between SEQ ID NO:2 of the invention and the SEQ ID NO:2 of '003 at page 16 is also very low (only 53%) (see Appendix B, part 3).

In summary, for the reasons set forth herein, Applicants maintain that claims 24-50 clearly and patentably define the invention. Applicants request that the Examiner reconsider and withdraw the various grounds for rejection set forth in the Office Action.

If the Examiner would like to discuss any of the issues raised in the Office Action, Applicants' representative can be reached at (617) 542-5070. Enclosed is a $1262 check for excess claim fees and a one month extension fee. Please apply any other charges or credits to deposit account 06-1050.

Respectfully submitted,

Reg No: 40,900

Janis K. Fraser, Ph.D., J.D. for
Reg. No. 34,819

Date: 12/16/03

Fish & Richardson P.C.
225 Franklin Street
Boston, MA 02110-2804
Telephone: (617) 542-5070
Facsimile: (617) 542-8906

10333584.doc

Deciphering the Message in Protein Sequences: Tolerance to Amino Acid Substitutions

JAMES U. BOWIE,* JOHN F. REIDHAAR-OLSON, WENDELL A. LIM, ROBERT T. SAUER

An amino acid sequence encodes a message that determines the shape and function of a protein. This message is highly degenerate in that many different sequences can code for proteins with essentially the same structure and activity. Comparison of different sequences with similar messages can reveal key features of the code and improve understanding of how a protein folds and how it performs its function.

THE GENOME IS MANIFEST LARGELY IN THE SET OF PROteins that it encodes. It is the ability of these proteins to fold into unique three-dimensional structures that allows them to function and carry out the instructions of the genome. Thus, comprehending the rules that relate amino acid sequence to structure is fundamental to an understanding of biological processes. Because an amino acid sequence contains all of the information necessary to determine the structure of a protein (*1*), it should be possible to predict structure from sequence, and subsequently to infer detailed aspects of function from the structure. However, both problems are extremely complex, and it seems unlikely that either will be solved in an exact manner in the near future. It may be possible to obtain approximate solutions by using experimental data to simplify the problem. In this article, we describe how an analysis of allowed amino acid substitutions in proteins can be used to reduce the complexity of sequences and reveal important aspects of structure and function.

Methods for Studying Tolerance to Sequence Variation

There are two main approaches to studying the tolerance of an amino acid sequence to change. The first method relies on the process of evolution, in which mutations are either accepted or rejected by natural selection. This method has been extremely powerful for proteins such as the globins or cytochromes, for which sequences from many different species are known (*2–7*). The second approach uses genetic methods to introduce amino acid changes at specific positions in a cloned gene and uses selections or screens to identify functional sequences. This approach has been used to great advantage for proteins that can be expressed in bacteria or yeast, where the appropriate genetic manipulations are possible (*3, 8–11*). The end results of both methods are lists of active sequences that can be compared and analyzed to identify sequence features that are essential for folding or function. If a particular property of a side chain, such as charge or size, is important at a given position, only side chains that have the required property will be allowed. Conversely, if the chemical identity of the side chain is unimportant, then many different substitutions will be permitted.

Studies in which these methods were used have revealed that proteins are surprisingly tolerant of amino acid substitutions (*2–4, 11*). For example, in studying the effects of approximately 1500 single amino acid substitutions at 142 positions in *lac* repressor, Miller and co-workers found that about one-half of all substitutions were phenotypically silent (*11*). At some positions, many different, nonconservative substitutions were allowed. Such residue positions play little or no role in structure and function. At other positions, no substitutions or only conservative substitutions were allowed. These residues are the most important for *lac* repressor activity.

What roles do invariant and conserved side chains play in proteins? Residues that are directly involved in protein functions such as binding or catalysis will certainly be among the most conserved. For example, replacing the Asp in the catalytic triad of trypsin with Asn results in a 10^4-fold reduction in activity (*12*). A similar loss of activity occurs in λ repressor when a DNA binding residue is changed from Asn to Asp (*13*). To carry out their function, however, these catalytic residues and binding residues must be precisely oriented in three dimensions. Consequently, mutations in residues that are required for structure formation or stability can also have dramatic effects on activity (*10, 14–16*). Hence, many of the residues that are conserved in sets of related sequences play structural roles.

Substitutions at Surface and Buried Positions

In their initial comparisons of the globin sequences, Perutz and co-workers found that most buried residues require nonpolar side chains, whereas few features of surface side chains are generally conserved (*6*). Similar results have been seen for a number of protein families (*2, 4, 5, 7, 17, 18*). An example of the sequence tolerance at surface versus buried sites can be seen in Fig. 1, which shows the allowed substitutions in λ repressor at residue positions that are near the dimer interface but distant from the DNA binding surface of the protein (*9*). These substitutions were identified by a functional

The authors are in the Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139.

*Present address: Department of Chemistry and Biochemistry and the Molecular Biology Institute, University of California, Los Angeles, Los Angeles, CA 90024.

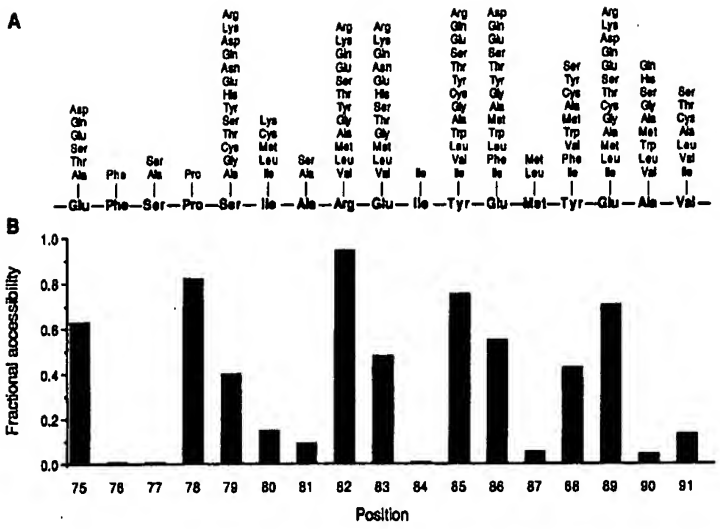


Fig. 1. (**A**) Amino acid substitutions allowed in a short region of λ repressor. The wild-type sequence is shown along the center line. The allowed substitutions shown above each position were identified by randomly mutating one to three codons at a time by using a cassette method and applying a functional selection (*9*). (**B**) The fractional solvent accessibility (*42*) of the wild-type side chain in the protein dimer (*43*) relative to the same atoms in an Ala-X-Ala model tripeptide.

selection after cassette mutagenesis. A histogram of side chain solvent accessibility in the crystal structure of the dimer is also shown in Fig. 1. At six positions, only the wild-type residue or relatively conservative substitutions are allowed. Five of these positions are buried in the protein. In contrast, most of the highly exposed positions tolerate a wide range of chemically different side chains, including hydrophilic and hydrophobic residues. Hence, it seems that most of the structural information in this region of the protein is carried by the residues that are solvent inaccessible.

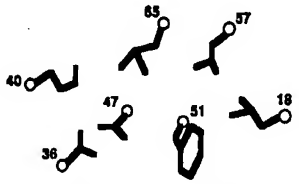
Constraints on Core Sequences

Because core residue positions appear to be extremely important for protein folding or stability, we must understand the factors that dictate whether a given core sequence will be acceptable. In general, only hydrophobic or neutral residues are tolerated at buried sites in proteins, undoubtedly because of the large favorable contribution of the hydrophobic effect to protein stability (*19*). For example, Fig. 2 shows the results of genetic studies used to investigate the substitutions allowed at residue positions that form the hydrophobic core of the NH_2-terminal domain of λ repressor (*20*). The acceptable core sequences are composed almost exclusively of Ala, Cys, Thr, Val, Ile, Leu, Met, and Phe. The acceptability of many different residues at each core position presumably reflects the fact that the hydrophobic effect, unlike hydrogen bonding, does not depend on specific residue pairings. Although it is possible to imagine a hypothetical core structure that is stabilized exclusively by residues forming hydrogen bonds and salt bridges, such a core would probably be difficult to construct because hydrogen bonds require pairing of donors and acceptors in an exact geometry. Thus the repertoire of possible structures that use a polar core would probably be extremely limited (*21*). Polar and charged residues are occasionally found in the cores of proteins, but only at positions where their hydrogen bonding needs can be satisfied (*22*).

The cores of most proteins are quite closely packed (*23*), but some volume changes are acceptable. In λ repressor, the overall core volume of acceptable sequences can vary by about 10%. Changes at individual sites, however, can be considerably larger. For example, as shown in Fig. 2, both Phe and Ala are allowed at the same core position in the appropriate sequence contexts. Large volume changes at individual buried sites have also been observed in phylogenetic studies, where it has been noted that the size decreases and increases at interacting residues are not necessarily related in a simple complementary fashion (*5, 7, 17*). Rather, local volume changes are accommodated by conformational changes in nearby side chains and by a variety of backbone movements.

The Informational Importance of the Core

With occasional exceptions, the core must remain hydrophobic and maintain a reasonable packing density. However, since the core is composed of side chains that can assume only a limited number of conformations (*24*), efficient packing must be maintained without steric clashes. How important are hydrophobicity, volume, and steric complementarity in determining whether a given sequence can form an acceptable core? Each factor is essential in a physical sense, as a stable core is probably unable to tolerate unsatisfied hydrogen bonding groups, large holes, or steric overlaps (*25*). However, in an informational sense, these factors are not equivalent. For example, in experiments in which three core residues of λ repressor were mutated simultaneously, volume was a relatively unimportant informational constraint because three-quarters of all possible combinations of the 20 naturally occurring amino acids had volumes within the range tolerated in the core, and yet most of these sequences were unacceptable (*20*). In contrast, of the sequences that contained only



40	36	47	65	51	57	18
Ala	Ala	Ala	Ala	Cys	Cys	Ala
Cys	Cys	Cys	Ser	Val	Pro	Cys
Val	Thr	Thr	Cys	Ile	Val	Val
Ile	Val	Val	Thr	Leu	Ile	Ile
Leu	Ile	Ile	Val	Met	Leu	Leu
Met	Leu	Leu	Ile	Phe	Met	Met
Phe	Met	Met	Leu		Phe	
	Phe	Phe	Met			
			Phe			

Fig. 2. Amino acid substitutions allowed in the core of λ repressor. The wild-type side chains are shown pictorially in the approximate orientation seen in the crystal structure (*43*). The lists of allowed substitutions at each position are shown below the wild-type side chains. These substitutions were identified by randomly mutating one to four residues at a time by using a cassette method and applying a functional selection (*20*). Not all substitutions are allowed in every sequence background.

the appropriate hydrophobic residues, a significant fraction were acceptable. Hence, the hydrophobicity of a sequence contains more information about its potential acceptability in the core than does the total side chain volume. Steric compatibility was intermediate between volume and hydrophobicity in informational importance.

The Informational Importance of Surface Sites

We have noted that many surface sites can tolerate a wide variety of side chains, including hydrophilic and hydrophobic residues. This result might be taken to indicate that surface positions contain little structural information. However, Bashford *et al.*, in an extensive analysis of globin sequences (*4*), found a strong bias against large hydrophobic residues at many surface positions. At one level, this may reflect constraints imposed by protein solubility, because large patches of hydrophobic surface residues would presumably lead to aggregation. At a more fundamental level, protein folding requires a partitioning between surface and buried positions. Consequently, to achieve a unique native state without significant competition from other conformations, it may be important that some sites have a decided preference for exterior rather than interior positions. As a result, many surface sites can accept hydrophobic residues individually, but the surface as a whole can probably tolerate only a moderate number of hydrophobic side chains.

Identification of Residue Roles from Sets of Sequences

Often, a protein of interest is a member of a family of related sequences. What can we infer from the pattern of allowed substitutions at positions in sets of aligned sequences generated by genetic or phylogenetic methods? Residue positions that can accept a number of different side chains, including charged and highly polar residues, are almost certain to be on the protein surface. Residue positions that remain hydrophobic, whether variable or not, are likely to be buried within the structure. In Fig. 3, those residue positions in λ repressor that can accept hydrophilic side chains are shown in orange and those that cannot accept hydrophilic side chains are shown in green. The obligate hydrophobic positions define the core of the structure, whereas positions that can accept hydrophilic side chains define the surface.

Functionally important residues should be conserved in sets of active sequences, but it is not possible to decide whether a side chain is functionally or structurally important just because it is invariant or conserved. To make this distinction requires an independent assay of protein folding. The ability of a mutant protein to maintain a stably folded structure can often be measured by biophysical techniques, by susceptibility to intracellular proteolysis (*26*), or by binding to antibodies specific for the native structure (*27, 28*). In the latter cases, it is possible to screen proteins in mutated clones for the ability to fold even if these proteins are inactive. Sets of sequences that allow formation of a stable structure can then be compared to the sets that allow both folding and function, with the active site or binding residues being those that are variable in the set of stable proteins but invariant in the set of functional proteins. The DNA-binding residues of Arc repressor were identified by this method (*8*). The receptor-binding residues of human growth hormone were also identified by comparing the stabilities and activities of a set of mutant sequences (*28*). However, in this case, the mutants were generated as hybrid sequences between growth hormone and related hormones with different binding specificities.

Implications for Structure Prediction

At present, the only reliable method for predicting a low-resolution tertiary structure of a new protein is by identifying sequence similarity to a protein whose structure is already known (*29, 30*). However, it is often difficult to align sequences as the level of sequence similarity decreases, and it is sometimes impossible to detect statistically significant sequence similarity between distantly related proteins. Because the number of known sequences is far greater than the number of known structures, it would be advantageous to increase the reach of the available structural information by improving methods for detecting distant sequence relations and for subsequently aligning these sequences based on structural principles. In a normal homology search, the sequence database is scanned with a single test sequence, and every residue must be weighted equally. However, some residues are more important than others and should be weighted accordingly. Moreover, certain regions of the protein are more likely to contain gaps than others. Both kinds of information can be obtained from sequence sets, and several techniques have

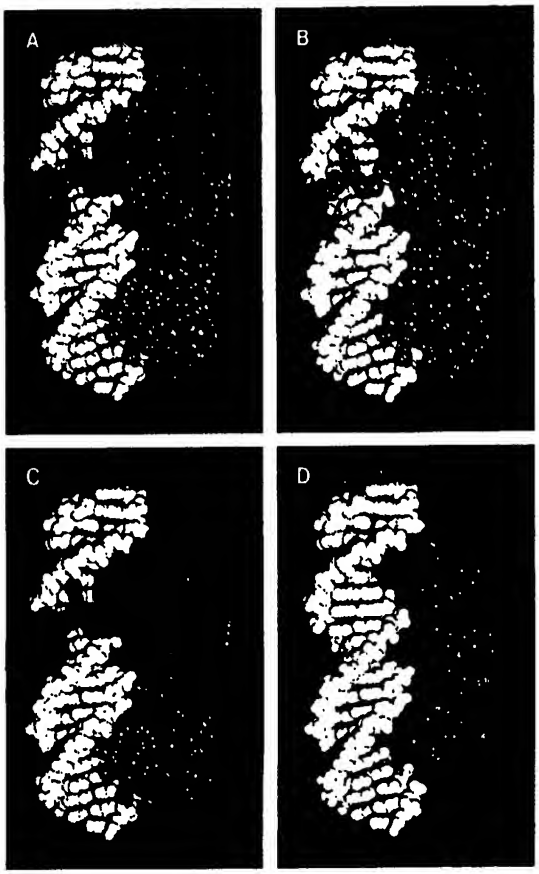


Fig. 3. Tolerance of positions in the NH_2-terminal domain of λ repressor to hydrophilic side chains. The complex (*43*) of the repressor dimer (blue) and operator DNA (white) is shown. In (**A**), positions that can tolerate hydrophilic side chains are shown in orange. The same side chains are shown in (**B**) without the remaining protein atoms. In (**C**), positions that require hydrophobic or neutral side chains are shown in green. These side chains are shown in (**D**) without the remaining protein atoms. About three-fourths of the 92 side chains in the NH_2-terminal domain are included in both (B) and (D). The remaining positions have not been tested. Data are from (*9, 14, 20, 27, 44*).

been used to combine such information into more appropriately weighted sequence searches and alignments (*31*). These methods were used to align the sequences of retroviral proteases with aspartic proteases, which in turn allowed construction of a three-dimensional model for the protease of human immunodeficiency virus type 1 (*29*). Comparison with the recently determined crystal structure of this protein revealed reasonable agreement in many areas of the predicted structure (*32*).

The structural information at most surface sites is highly degenerate. Except for functionally important residues, exterior positions seem to be important chiefly in maintaining a reasonably polar surface. The information contained in buried residues is also degenerate, the main requirement being that these residues remain hydrophobic. Thus, at its most basic level, the key structural message in an amino acid sequence may reside in its specific pattern of hydrophobic and hydrophilic residues. This is meant in an informational sense. Clearly, the precise structure and stability of a protein depends on a large number of detailed interactions. It is possible, however, that structural prediction at a more primitive level can be accomplished by concentrating on the most basic informational aspects of an amino acid sequence. For example, amphipathic patterns can be extracted from aligned sets of sequences and used, in some cases, to identify secondary structures.

If a region of secondary structure is packed against the hydrophobic core, a pattern of hydrophobic residues reflecting the periodicity of the secondary structure is expected (*33, 34*). These patterns can be obscured in individual sequences by hydrophobic residues on the protein surface. It is rare, however, for a surface position to remain hydrophobic over the course of evolution. Consequently, the amphipathic patterns expected for simple secondary structures can be much clearer in a set of related sequences (*6*). This principle is illustrated in Fig. 4, which shows helical hydrophobic moment plots for the Antennapedia homeodomain sequence (Fig. 4A) and for a composite sequence derived from a set of homologous homeodomain proteins (Fig. 4B) (*35*). The hydrophobic moment is a simple measure of the degree of amphipathic character of a sequence in a given secondary structure (*34*). The amphipathic character of the three α-helical regions in the Antennapedia protein (*36*) is clearly revealed only by the analysis of the combined set of homeodomain sequences. The secondary structure of Arc repressor, a small DNA-binding protein, was recently predicted by a similar method (*8*) and confirmed by nuclear magnetic resonance studies (*37*).

The specific pattern of hydrophobic and hydrophilic residues in an amino acid sequence must limit the number of different structures a given sequence can adopt and may indeed define its overall fold. If this is true, then the arrangement of hydrophobic and hydrophilic residues should be a characteristic feature of a particular fold. Sweet and Eisenberg have shown that the correlation of the pattern of hydrophobicity between two protein sequences is a good criterion for their structural relatedness (*38*). In addition, several studies indicate that patterns of obligatory hydrophobic positions identified from aligned sequences are distinctive features of sequences that adopt the same structure (*4, 29, 38, 39*). Thus, the order of hydrophobic and hydrophilic residues in a sequence may actually be sufficient information to determine the basic folding pattern of a protein sequence.

Although the pattern of sequence hydrophobicity may be a characteristic feature of a particular fold, it is not yet clear how such patterns could be used for prediction of structure de novo. It is important to understand how patterns in sequence space can be related to structures in conformation space. Lau and Dill have approached this problem by studying the properties of simple sequences composed only of H (hydrophobic) and P (polar) groups on two-dimensional lattices (*40*). An example of such a representation is shown in Fig. 5. Residues adjacent in the sequence must occupy adjacent squares on the lattice, and two residues cannot occupy the same space. Free energies of particular conformations are evaluated with a single term, an attraction of H groups. By considering chains of ten residues, an exhaustive conformational search for all 1024 possible sequences of H and P residues was possible. For longer sequences only a representative fraction of the allowed sequence or conformation space could be explored. The significant results were as follows: (i) not all sequences can fold into a "native" structure and only a few sequences form a unique native structure; (ii) the probability that a sequence will adopt a unique native structure increases with chain length; and (iii) the native states are compact, contain a hydrophobic core surrounded by polar residues, and contain significant secondary structure. Although the gap between these two-dimensional simulations and three-dimensional structures is large, the use of simple rules and sequence representations yields results similar to those expected for real proteins. Three-dimensional lattice methods are also beginning to be developed and evaluated (*41*).

Summary

There is more information in a set of related sequences than in a single sequence. A number of practical applications arise from an analysis of the tolerance of residue positions to change. First, such information permits the evaluation of a residue's importance to the function and stability of a protein. This ability to identify the essential elements of a protein sequence may improve our understanding of the determinants of protein folding and stability as well as protein function. Second, patterns of tolerance to amino acid substitutions of varying hydrophilicity can help to identify residues likely to be buried in a protein structure and those likely to occupy

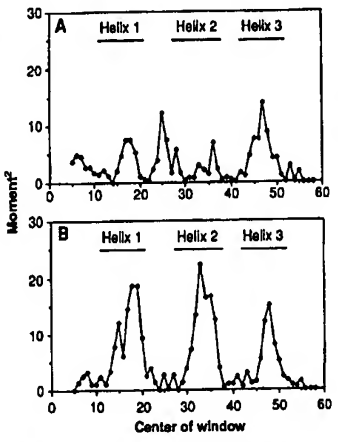


Fig. 4. Helical hydrophobic moments calculated by using (**A**) the Antennapedia homeodomain sequence or (**B**) a set of 39 aligned homeodomain sequences (*35*). The bars indicate the extent of the helical regions identified in nuclear magnetic resonance studies of the Antennapedia homeodomain (*36*). To determine hydrophobic moments, residues were assigned to one of three groups: H1 (high hydrophobicity = Trp, Ile, Phe, Leu, Met, Val, or Cys); H2 (medium hydrophobicity = Tyr, Pro, Ala, Thr, His, Gly, or Ser); and H3 (low hydrophobicity = Gln, Asn, Glu, Asp, Lys, or Arg). For the aligned homeodomain sequences, the residues at each position were sorted by their hydrophobicity by using the scale of Fauchere and Pliska (*45*). Arg and Lys were not counted unless no other residue was found at the position, because they contain long aliphatic side chains and can thereby substitute for nonpolar residues at some buried sites. To account for possible sequence errors and rare exceptions, the most hydrophilic residue allowed at each position was discarded unless it was observed twice. The second most hydrophilic residue was then chosen to represent the hydrophobicity of each position. An eight-residue window was used and the vectors projected radially every 100°. The vector magnitudes were assigned a value of 1, 0, or −1 for positions where the hydrophobicity group was H1, H2, or H3, respectively.

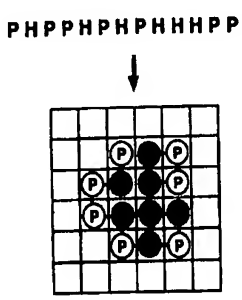


Fig. 5. A representation of one compact conformation for a particular sequence of H and P residues on a two-dimensional square lattice. [Adapted from (*40*), with permission of the American Chemical Society]

surface positions. The amphipathic patterns that emerge can be used to identify probable regions of secondary structure. Third, incorporating a knowledge of allowed substitutions can improve the ability to detect and align distantly related proteins because the essential residues can be given prominence in the alignment scoring.

As more sequences are determined, it becomes increasingly likely that a protein of interest is a member of a family of related sequences. If this is not the case, it is now possible to use genetic methods to generate lists of allowed amino acid substitutions. Consequently, at least in the short term, it may not be necessary to solve the folding problem for individual protein sequences. Instead, information from sequence sets could be used. Perhaps by simplifying sequence space through the identification of key residues, and by simplifying conformation space as in the lattice methods, it will be possible to develop algorithms to generate a limited number of trial structures. These trial structures could then, in turn, be evaluated by further experiments and more sophisticated energy calculations.

REFERENCES AND NOTES

1. C. J. Epstein, R. F. Goldberger, C. B. Anfinsen, *Cold Spring Harbor Symp. Quant. Biol.* **28**, 439 (1963); C. B. Anfinsen, *Science* **181**, 223 (1973).
2. R. E. Dickerson, *Sci. Am.* **242**, 136 (March 1980).
3. M. D. Hampsey, G. Das, F. Sherman, *FEBS Lett.* **231**, 275 (1988).
4. D. Bashford, C. Chothia, A. M. Lesk, *J. Mol. Biol.* **196**, 199 (1987).
5. A. M. Lesk and C. Chothia, *ibid.* **136**, 225 (1980).
6. M. F. Perutz, J. C. Kendrew, H. C. Watson, *ibid.* **13**, 669 (1965).
7. C. Chothia and A. M. Lesk, *Cold Spring Harbor Symp. Quant. Biol.* **52**, 399 (1965).
8. J. U. Bowie and R. T. Sauer, *Proc. Natl. Acad. Sci. U.S.A.* **86**, 2152 (1989).
9. J. F. Reidhaar-Olson and R. T. Sauer, *Science* **241**, 53 (1988); *Proteins Struct. Funct. Genet.*, in press.
10. D. Shortle, *J. Biol. Chem.* **264**, 5315 (1989).
11. J. H. Miller *et al.*, *J. Mol. Biol.* **131**, 191 (1979).
12. S. Sprang *et al.*, *Science* **237**, 905 (1987); C. S. Craik, S. Roczniak, C. Largman, W. J. Rutter, *ibid.*, p. 909.
13. H. C. M. Nelson and R. T. Sauer, *J. Mol. Biol.* **192**, 27 (1986).
14. M. H. Hecht, J. M. Sturtevant, R. T. Sauer, *Proc. Natl. Acad. Sci. U.S.A.* **81**, 5685 (1984).
15. T. Alber, D. Sun, J. A. Nye, D. C. Muchmore, B. W. Matthews, *Biochemistry* **26**, 3754 (1987).
16. D. Shortle and A. K. Meeker, *Proteins Struct. Funct. Genet.* **1**, 81 (1986).
17. A. M. Lesk and C. Chothia, *J. Mol. Biol.* **160**, 325 (1982).
18. W. R. Taylor, *ibid.* **188**, 233 (1986).
19. W. Kauzmann, *Adv. Protein Chem.* **14**, 1 (1959); R. L. Baldwin, *Proc. Natl. Acad. Sci. U.S.A.* **83**, 8069 (1986).
20. W. A. Lim and R. T. Sauer, *Nature* **339**, 31 (1989); in preparation.
21. Lesk and Chothia (*5*) have argued that a protein core composed solely of hydrogen-bonded residues would also be inviable on evolutionary grounds, as a mutational change in one core residue would require compensating changes in any interacting residue or residues to maintain a stable structure.
22. T. M. Gray and B. W. Matthews, *J. Mol. Biol.* **175**, 75 (1984); E. N. Baker and R. E. Hubbard, *Prog. Biophys. Mol. Biol.* **44**, 97 (1984).
23. F. M. Richards, *J. Mol. Biol.* **82**, 1 (1974).
24. J. W. Ponder and F. M. Richards, *ibid.* **193**, 775 (1987).
25. J. T. Kellis, Jr., K. Nyberg, A. R. Fersht, *Biochemistry* **28**, 4914 (1989); W. S. Sandberg and T. C. Terwilliger, *Science* **245**, 54 (1989).
26. A. A. Pakula and R. T. Sauer, *Proteins Struct. Funct. Genet.* **5**, 202 (1989).
27. B. C. Cunningham and J. A. Wells, *Science* **244**, 1081 (1989); R. M. Breyer and R. T. Sauer, *J. Biol. Chem.* **264**, 13348 (1989).
28. B. C. Cunningham, P. Jhurani, P. Ng, J. A. Wells, *Science* **243**, 1330 (1989).
29. L. H. Pearl and W. R. Taylor, *Nature* **329**, 351 (1987).
30. W. J. Brown *et al.*, *J. Mol. Biol.* **42**, 65 (1969); J. Greer, *ibid.* **153**, 1027 (1981); J. M. Berg, *Proc. Natl. Acad. Sci. U.S.A.* **85**, 99 (1988).
31. W. R. Taylor, *Protein Eng.* **2**, 77 (1988).
32. M. A. Navia *et al.*, *Nature* **337**, 615 (1989).
33. M. Schiffer and A. B. Edmundson, *Biophys. J.* **7**, 121 (1967); V. I. Lim, *J. Mol. Biol.* **88**, 857 (1974); *ibid.*, p. 873.
34. D. Eisenberg, R. M. Weiss, T. C. Terwilliger, *Nature* **299**, 371 (1982); D. Eisenberg, D. Schwarz, M. Komaromy, R. Wall, *J. Mol. Biol.* **179**, 125 (1984); D. Eisenberg, R. M. Weiss, T. C. Terwilliger, *Proc. Natl. Acad. Sci. U.S.A.* **81**, 140 (1984).
35. T. R. Burglin, *Cell* **53**, 339 (1988).
36. G. Otting *et al.*, *EMBO J.* **7**, 4305 (1988).
37. J. N. Breg, R. Boelens, A. V. E. George, R. Kaptein, *Biochemistry* **28**, 9826 (1989); M. G. Zagorski, J. U. Bowie, A. K. Vershon, R. T. Sauer, D. J. Patel, *ibid.*, p. 9813.
38. R. M. Sweet and D. Eisenberg, *J. Mol. Biol.* **171**, 479 (1983).
39. J. U. Bowie, N. D. Clarke, C. O. Pabo, R. T. Sauer, *Proteins Struct. Funct. Genet.*, in preparation.
40. K. F. Lau and K. A. Dill, *Macromolecules* **22**, 3986 (1989).
41. A. Sikorski and J. Skolnick, *Proc. Natl. Acad. Sci. U.S.A.* **86**, 2668 (1989); A. Kolinski, J. Skolnick, R. Yaris, *Biopolymers* **26**, 937 (1987); D. G. Covell and R. L. Jernigan, *Biochemistry*, in press.
42. B. Lee and F. M. Richards, *J. Mol. Biol.* **55**, 379 (1971).
43. S. R. Jordan and C. O. Pabo, *Science* **242**, 893 (1988).
44. R. M. Breyer, thesis, Massachusetts Institute of Technology, Cambridge (1988).
45. J.-L. Fauchere and V. Pliska, *Eur. J. Med. Chem.-Chim. Ther.* **18**, 369 (1983).

46. We thank C. O. Pabo and S. Jordan for coordinates of the NH_2-terminal domain of λ repressor and its operator complex. We also thank P. Schimmel for the use of his graphics system and J. Burnbaum and C. Francklyn for assistance. Supported in part by NIH grant AI-15706 and predoctoral grants from NSF (J.R.-O.) and Howard Hughes Medical Institute (W.A.L.).

Appendix B

1. Homology between SEQ ID NO:2 of the invention and the amino acid sequen

Score = 358 bits (920), Expect = 1e-97
Identities = 184/378 (48%), Positives = 249/378 (65%), Gaps = 3/378 (0%)

Query : SEQ ID NO:2 of the invention
Sbjct : YALOGOW of D1

```
Query: 1   MKGLLYYGTNDIRYSETVPEPEIKNPNDVKIKVSYCGICGTDLKEFTYSGGPVFFPKQGT 60
           M+ L Y+   DI ++  +P PEI+  ++V I VS+CGICG+DL E  Y  GP+F PK G
Sbjct: 1   MRALAYFKKGDIHFTNDIPRPEIQTDDEVIIDVSWCGICGSDLHE---YLDGPIFMPKDGE 58

Query: 61  KDKISGYELPLCPGHEFSGTVVEVGSGVTSVKPGDRVAVEATSHCSDRSRYKDTVAQDLG 120
             K+S   LPL  GHE SG V +VG  VT VK GD V V+A S C+D  + +  +
Sbjct: 59  CMKLSNAALPLAMGHEMSGIVSKVGPKVTKVKVGDHVVVDAASSCADLHCWPHSKFYNSK 118

Query: 121 LCMACQSGSPNCCASLSFCGLGGASGGFAEYVVYGEDHMVKLPDSIPDDIGALVEPISVA 180
            C ACQ GS N C    F GLG  SGGFAE VV  + H++ +P  IP D+ ALVEP+SV
Sbjct: 119 PCDACQRGSENLCTHAGFVGLGVISGGFAEQVVVSQHHIIPVPKEIPLDVAALVEPLSVT 178

Query: 181 WHAVERARFQPGQTALVLGGGPIGLATILALQGHHAGKIVCSEPALIRRQFAKELGAEVF 240
           WHAV+ + F+ Q +ALVLG GPIGL TIL L+G  A KIV SE A  R + AK+LG EVF
Sbjct: 179 WHAVKISGFKKGSSALVLGAGPIGLCTILVLKGMGASKIVVSEIAERRIEMAKKLGVEVF 238

Query: 241 DPSTCDDAN-AVLKAMVPENEGFHAAFDCSGVPQTFTTSIVATGPSGIAVNVAVWGDHPI 299
           +PS     + +L+ +   ++GF  ++DCSG+  TF TS+ A   G A N+AVWG  P+
Sbjct: 239 NPSKHGHKSIEILRGLTKSHDGFDYSYDCSGIQVTFETSLKALTFKGTATNIAVWGPKPV 298

Query: 300 GFMPMSLTYQEKYATGSMCYTVKDFQEVVKALEDGLISLDKARKMITGKVHLKDGVEKGF 359
            F PM +T QEK  TGS+ Y V+ F+EVV+A+ +G I+++  ++ITGK  +DG EKGF
Sbjct: 299 PFQPMDVTLQEKVMTGSIGYVVEAFEEVVRAIHNGDIAMEDCKQLITGKQRIEDGWEKGF 358

Query: 360 KQLIEHKENNVKILVTPN 377
           ++L++HKE+NVKIL+TPN
Sbjct: 359 QELMDHKESNVKILLTPN 376
```

2. Homology between SEQ ID NO:1 of the invention and the SEQ ID NO: 1 of D3

Query : SEQ ID NO:1 of the Invention (1143 letters)
Sbjct = SEC UD NO:1 of D3 (Length = 1168)

Score = 24 bits (12), Expect = 0.064
Identities = 12/12 (100%)

Strand = Plus / Plus

Query: 718 ttcgatccttct 729
||||||||||||
Sbjct: 712 ttcgatccttct 723

Score = 24 bits (12), Expect = 0.064
Identities = 12/12 (100%)
Strand = Plus / Plus

Query: 541 tggcatgctgtt 552
||||||||||||
Sbjct: 535 tggcatgctgtt 546

Score = 22 bits (11), Expect = 0.26
Identities = 11/11 (100%)
Strand = Plus / Plus

Query: 479 tcaagctgcca 489
|||||||||||
Sbjct: 684 tcaagctgcca 694

Score = 22 bits (11), Expect = 0.25
Identities = 11/11 (100%)
Strand = Plus / Plus

Query: 892 ccaattggatt 902
|||||||||||
Sbjct: 598 ccaattggatt 608

3. Homology between SEQ ID NO:2 of the invention and the SEQ ID NO: 2 of D3

Query : SEQ ID NO:2 of the invention (380 letters)
Sbjct = SEQ ID NO:2 of D3 (Length = 400)

Score = 396 bits (1018), Expect = e-115
Identities = 201/378 (53%), Positives = 254/378 (67%), Gaps = 3/378 (0%)

```
Query: 1   MKGLLYYGTNDIRYSETVPEPEIKNPNDVKIKVSYCGICGTDLKEFTYSGGPVFFPKQGT 60
           M+ L Y+G  DIRY++ + EP I+  + ++I+VS+CGICG+DL E  Y  GP+FFP+ G
Sbjct: 1   MRALAYFGKQDIRYTKDLEEPVIETDDGIEIEVSWCGICGSDLHE---YLDGPIFFPEDGK 58

Query: 61  KDKISGYELPLCPGHEFXXXXXXXXXXXXXXKPGDRVAVEATSHCSDRSRYKDTVAQDLG 120
              +SG  LP   GHE                K GD V VEAT  C D   + +
Sbjct: 59  VHDVSGLGLPQAMGHEMSGIVSKVGPKVTNIKAGDHVVVEATGTCLDHYTWPNAAHAKDA 118

Query: 121 LCMACQSGSPNCCASLSFCGLGQASGGFAEYVVYGEDHMVKLPDSIPDDIGALVEPISVA 180
            C ACQ G  NCCA L F GLG  SGGFAE VV  E H+VK+P+++P D+ ALVEPISV+
Sbjct: 119 ECAACQRGFYNCCAHLGFMGLGVHSGGFAEKVVVSEKHVVKIPNTLPLDVAALVEPISVS 178

Query: 181 WHAVERARFQPGQTALVLGGGPIGLATILALQGHHAGKIVCSEPALIRRQFAKELGAEVF 240
           WHAV  ++ Q GQ+ALVLG GPIGLATILALQGH A KIV SEPA IRR  A +LG E F
Sbjct: 179 WHAVRISKLQKGQSALVLGAGPIGLATILALQGHGASKIVVSEPAEIRRNQAAKLGVETF 238

Query: 241 DPST-CDDANAVLKAMVPENEGFHAAFDCSGVPQTFTTSIVATGPSGIAVNVAVWGDHPI 299
           DPS   +DA  +LK + P  EGF  A+DCSGV  TF T + AT   G+ VN+A+WG  PI
Sbjct: 239 DPSEHKEDAVNILKKLAPGGEGFDFAYDCSGVKPTFDTGVHATTFRGMYVNIAIWGHKPI 298

Query: 300 GFMPMSLTYQEKYATGSMCYTVKDFQEVVKALEDGLISLDKARKMITGKVHLKDGVEKGF 359
            F PM +T QEK+ TGSMCYT+KDF++VV+AL +G I++DKAR +ITG+  ++DG  KGF
Sbjct: 299 DFKPMDVTLQEKFVTGSMCYTIKDFEDVVQALGNGSIAMDKARHLITGRQKIEDGFTKGF 358

Query: 360 KQLIEHKENNVKILVTPN 377
           +L+ HKE N+KIL+TPN
Sbjct: 359 DELMNHKEKNIKILLTPN 376
```